Econometría Fácil

Una guía simple a las complejidades del análisis de regresión

Matías Cabello

Edición preliminar.[1]
Fecha: 7 de marzo de 2016

Atención: Si no imprime en formato duplex (por ambas caras), procure nunca ser descubierto por el autor de este libro. Cuidemos al planeta ;)

[1] Agradezco a Mathias Legrand por su maravilloso *Legrand Orange Book template*, el cual he utilizado para redactar el libro. También agradezco a la comunidad de `stackexchange.com` por resolver tantas dudas de programación de manera desinteresada. Agradezco a los desarrolladores de LATEX y a los desarrolladores de GRETL por entregarnos herramientas tan valiosas para la docencia e investigación sin ningún interés económico. Como no, también agradezco a mis alumnos y ayudantes por comentarios y correcciones a ediciones anteriores.

Introducción

La econometría (textualmente "medición económica") corresponde a un conjunto de métodos estadísticos empleados para analizar los fenómenos que estudian los economistas. ¿Cuáles son estos fenómenos? Podríamos pensar que son problemas microeconómicos relacionados con la empresa, los consumidores y los mercados, o con problemas de la macroeconomía como los ciclos económicos, el desempleo, la inflación, la inequidad y la eliminación de la pobreza. Si bien esta noción es correcta, la verdad es que los economistas claramente ya se aburrieron de estudiar estos temas y han expandido el abanico para incluir una serie de curiosidades relacionados con medio ambiente, psicología, cultura, salud, felicidad, racismo, conflictos diplomáticos y relaciones de pareja, por dar un par de ejemplos. Es decir, hoy la economía como disciplina es, en las palabras irónicas del economista surcoreano Ha-Joon Chang, "la explicación definitiva de la vida, el universo y de todo lo que existe".[2] Aunque esta última descripción es un tanto exagerada, el punto es que son pocos los temas que no se estudien dentro de la economía.

¿Cómo se explica que una disciplina originalmente acotada se transformara en un campo de estudio tan extensivo? Parte de la respuesta se encuentra en la enorme versatilidad de las herramientas que utilizan los economistas, dentro de las cuales se encuentran, sin duda, la técnicas econométricas que se presentan en este libro.

¿De qué se trata, exactamente, la econometría? Al menos en un libro introductorio como este, la econometría es sinónimo del **análisis de regresión múltiple**. La idea básica sobre la cuál se construye dicho análisis es que en el mundo existen procesos (denotemos a uno de ellos por y), los cuales son consecuencia de un número de causales (llamemos a dos de ellas x y z). Matemáticamente:

$$y = f(x, z, \ldots) \tag{1}$$

Es decir, y es una función de x, z y otras variables explicativas. ¿Cómo cambia y si aumentamos x, o si disminuimos z, o si tanto x como z toman conjuntamente los valores 3 y 5.5 respectivamente? Este es el tipo de preguntas que intenta responder la econometría mediante el análisis de regresión. La idea básica es tan general que se aplica a un sinnúmero de problemas prácticos. He a continuación un par de ejemplos.

Ganar dinero: En los negocios es importante saber cuál es la curva de demanda que tienen los productos ofertados. Si aumento el precio (x) del yoghurt en 10 %,

[2] Esta es una frase que Chang a repetido en varias conferencias. Claramente se encuentra parafraseando a "The Hitchhiker's Guide to the Galaxy" de Douglas Adam.

pero aumento también la visibilidad del producto (z) de 3 a 6, ¿aumenta o disminuye la cantidad vendida (y)? ¿En cuánto? ¿Qué tan seguro estoy de mis resultados? Las técnicas econométricas sirven para abordar este tipo de interrogantes. Por este motivo el mercado laboral ve con muy buenos ojos a profesionales con manejo avanzado de estas técnicas. Los bancos, por ejemplo, deben decidir a quién ofrecer sus créditos y bajo qué condiciones. Para ello deben calcular el perfil de riesgo (y) del cliente, en función de diversas variables explicativas ($x, z, \ldots$) como el nivel de ingreso, la edad, el grado académico, el número de hijos, etc. Los ejecutivos del banco muchas veces se rigen por un cálculo "mágico" que les entrega un computador. Bien, quien diseñó ese cálculo mágico es un econometrista (y si sabe hacer bien su trabajo puede ganar mucho, mucho dinero).

Para hacer análisis de regresión es fundamental tener acceso a datos. Mientras más datos hayan a disposición, más valioso es el conocimiento de las herramientas estadísticas para trabajar con ellos. Como nos encontramos en la era de la información y los datos guardados por compañías e instituciones crecen de manera exponencial, es fácil imaginar que el dominio de la econometría irá ganando importancia para poder hacer negocios exitosos.

Para ilustrar la prevalencia de la econometría en nuestras vidas, piense que cada vez que a usted le piden su número de identificación en un supermercado, o le ofrecen pagar con una tarjeta de la tienda para obtener un descuento, sus transacciones quedan registradas en la empresa. ¿Con qué objetivo? Si se tienen los conocimientos econométricos indicados, la información registrada sirve para explorar muchas interrogantes. Por dar un ejemplo, si y es el consumo de café de una cliente de supermercado (de la cual tenemos registrado cuál ha sido su patrón de comportamiento en el pasado, al igual que el de clientes similares) podríamos estudiar el valor esperado que debiera tomar su consumo (y) si, digamos, le gusta el chocolate (x) y se publicita el café con la imagen un hombre semi desnudo tomando café con un chocolate en la cama (z). No hay que ser un economista para saber que la publicidad aumenta las ventas. La pregunta de oro, que sólo se puede responder con herramientas estadísticas adecuadas y con un sabio uso de ellas, es en cuánto.

Comprender el mundo: Así como el estudio de las funciones del tipo (1) sirven para aumentar las ventas y los ingresos de una empresa, también sirven para entender una serie de fenómenos que nos rodean. Una pregunta abordada recientemente en una prestigiosa revista económica es: ¿Por qué algunos países son más machistas que otros? Hoy se sabe que las sociedades primitivas (al contrario de lo que muestran las películas) no eran machistas, sino igualitarias o incluso matriarcales. El estudio econométrico reveló que culturas que experimentaron antes la adopción del caballo y del arado desarrollaron una diferenciación de sexo mayor y hoy son sociedades con menor participación de la mujer en el mundo laboral, político y administrativo.[3]

¿Por qué son algunas personas más felices que otras? ¿Por qué existe el racismo en la mente de unos y no de otros? ¿Qué hay detrás de los gustos de las personas? ¿Qué

[3] Para quien se interese en el estudio: Alberto Alesina, Paola Giuliano, Nathan Nunn, 2013. "On the Origins of Gender Roles: Women and the Plough", *The Quarterly Journal of Economics*, vol. 128(2), pages 469-530.

hace que algunos países sean ricos y otros pobres? Todas estas preguntas y muchas otras pueden ser abordadas con las técnicas utilizadas en la econometría.

Pronosticar: Si conocemos la forma funcional $f(\cdot)$ que da origen a y, entonces basta con conocer qué valores tomarán las variables explicativas $x,z,\ldots$ para saber qué valor tomará y. Por ejemplo, si y corresponde a los milímetros cúbicos de lluvia caídos hoy, mientras $x,z,\ldots$ corresponden al conjunto de variables explicativas medidas ayer (presión atmosférica, temperatura del océano, humedad, etc.), entonces conocer la relación $f(\cdot)$ nos permite pronosticar la lluvia de *mañana* en función de las variables explicativas de hoy.

Modelos similares pueden ser aplicados para predecir las fluctuaciones cíclicas de la economía, el numero de clientes que tendrá una empresa el próximo año, el candidato presidencial que será elegido en un país o si un deudor será capaz de pagar sus deudas. En el peor de los casos y al estilo de las novelas de George Orwell, grupos de poder podrían hacer uso de cómo los celulares y computadores graban todo lo que hacemos, con quien nos relacionamos y cómo pensamos. A la luz del alcance de estas herramientas, cabe preguntarse: ¿cuánta información sobre nuestra vida privada debemos permitir recopilar a los servicios de inteligencia o conglomerados económicos, si esta información puede ser utilizada para pronosticar nuestro comportamiento, para aplacar movimientos políticos que pudieran afectar a los intereses establecidos, o simplemente ser vendida a empresas que hacen uso comercial de nuestra vida privada? Naturalmente, las predicciones son imprecisas, pero mejoran su calidad a medida que tenemos más datos para realizarlas, y vaya que crecen las bases de datos en estos días.

Ayudar al mundo: Por fortuna las técnicas empleadas en la econometría no sólo se utilizan para aumentar utilidades, hipnotizar a los clientes o mantener control total de la población al estilo orweliano. Muchos avances en ciencia y tecnología tienen su base en el análisis de regresión múltiple. Dentro de las aplicaciones más destacables cabe mencionar a la medicina. ¿Cuál es la efectividad, por ejemplo, de un medicamento para la prevención de infartos cardiacos? La probabilidad de tener un infarto podría ser y, la dosis del medicamento x y otras características del paciente z. Se podría estudiar la relación lineal

$$y = f(x,z,\ldots) = \beta_1 + \beta_2 x + \beta_3 z \tag{2}$$

Si β_2 es negativo, entonces el medicamento surge efecto. La magnitud de β_2 nos interesa para saber cuán efectivo es el medicamento. También será de gran interés la seguridad que se tiene respecto del valor estimado de β_2, tema que será tratado con profundidad a lo largo del texto.

De forma similar podemos diseñar muchas otras preguntas de utilidad para la humanidad ¿Cuál es el efecto de un programa de capacitación para la reducción de la extrema pobreza? ¿Cuál es el efecto de suplementar la dieta de un lactante sobre su desarrollo cerebral? ¿Cuáles son las claves para evitar las guerras, el egoísmo y fanatismo de las masas? La idea de este libro es presentar las herramientas estadísticas necesarias para abordar todas estas preguntas.

Capítulo 1

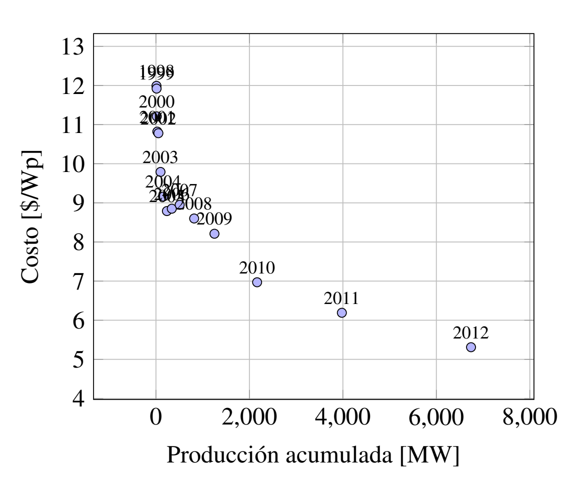
Mínimos cuadrados ordinarios (MCO)

1.1 ¿Cuándo será económicamente viable la energía solar?

Hoy el mundo se abastece principalmente de fuentes de energía fósil, no renovable y altamente contaminante. Alternativas renovables de energía con una baja huella de carbono, como la energía solar fotovoltaica, tienen un uso limitado debido a un alto precio de producción y ciertamente existen únicamente debido al aporte cuantioso de recursos por parte de un número limitado de estados (principalmente en Europa). Para poder competir de forma independiente con fuentes tradicionales, el costo de la energía fotovoltaica no debiera estar por sobre, digamos, 1 US$/watt-peak. ¿Será algún día esta tecnología competitiva en el mercado de energía?

En el desarrollo de toda tecnología observamos lo que se denomina técnicamente una “curva de aprendizaje”: a medida que más uso se hace de ella, más eficiente se vuelve. Los paneles fotovoltaicos no son la excepción. La figura 1.1 muestra la evolución del precio medio por watt en EE.UU. desde 1998, con el costo de un watt en el eje vertical y la cantidad acumulada de watts producidos desde los inicios de la implementación de la tecnología en el país. Resulta evidente que a mayor cantidad de Watts producidos baja el costo de producción. Es decir, existe una **correlación** negativa entre amabas variables.

Figura 1.1: Precio vs. experiencia en energía solar



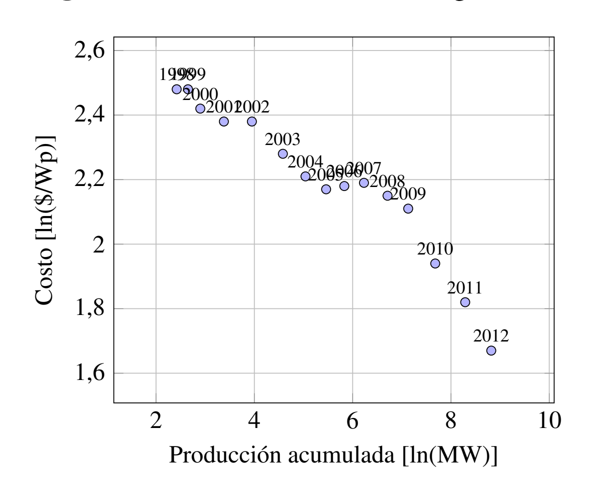
La figura 1.2 permite ver esta relación de forma más clara. En él se presenta el logaritmo del costo en el eje vertical y el logaritmo de la producción acumulada en el eje horizontal, los cuales se relacionan de una forma aproximadamente lineal.

Podríamos aproximar esta última relación negativa con una función del tipo

$$\underset{\text{Costo (log)}}{y} = \beta_1 + \beta_2 \underset{\text{Producción (log)}}{x} + \underset{\text{Otros factores}}{u}, \tag{1.1}$$

donde el logaritmo de la producción acumulada es la **variable explicativa** o **independiente** (la llamaremos simplemente x), y el logaritmo del costo por watt es la **variable explicada** o **dependiente** (la cual denotaremos con y).

Figura 1.2: Relación lineal en logaritmos



Aquello que se escapa a la relación lineal entre x e y se captura en u, denominado **error**. Como vemos en la figura, en torno al año 2006 el precio de la energía vio un aumento transitorio, el cual se debió a un alza en el precio del polisilicio, una materia prima fundamental para la elaboración de paneles solares. En la simplificación de la realidad que supone (1.1) este fenómeno se considera una de las muchas variables que podrían entrar en el error u. Si el precio del policilicio fuera tan importante explicando el precio como lo es el aprendizaje a lo largo de la curva, entonces tendría poco sentido tratarlo como error y en lugar de eso debiera figurar como otra variable explicativa. Pero si el efecto se neutraliza en el tiempo, entonces sí puede tener sentido asumir que se trata de un "error" e incluso podríamos asumir que este tiene valor esperado de cero sobre todas las unidades i analizadas,

$$\mathrm{E}[u_i] = 0. \tag{1.2}$$

Es decir, cada año, que denotaremos con i, se espera ex ante que todos los factores que se incluyen en el error sean cero (aunque ex post u_i siempre será negativo o positivo).

Suponer que el error tiene media cero nos sirve para hacer pronósticos, pues considerando (1.2) y suponiendo que está en nuestras manos (o de los estados financistas) decidir cuánto se invertirá en energía solar, tenemos

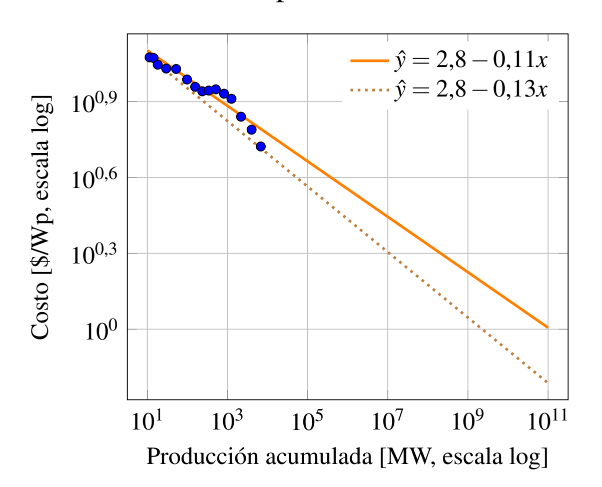
$$\mathrm{E}[y_i] = \beta_1 + \beta_2 x_i. \tag{1.3}$$

En palabras, el valor esperado del precio de energía solar, por ejemplo en el año 2018, depende sólo de cuánta energía se decidió producir con paneles fotovoltaicos hasta esa fecha. Así, dependiendo de cuáles sean los valores de β_1 y β_2 se puede indicar cuánto esfuerzo hay que destinar a la producción de energía solar para que esta sea finalmente competitiva.

La importancia de β_1 y β_2 queda clara en la figura 1.3. Ahí se muestra que para alcanzar una precio competitivo de 1 US$ por watt-peak se necesitan ya sea aproximadamente 1 millón de megawatt (10^9) o 100 millones de megawatt (10^{11}), ¡una suma bastante distinta!

En definitiva, nuestra función de **pronóstico** (o función de **ajuste** $\hat{y}$) va a depender de qué valores estimemos para los coeficientes de (1.1). Denotaremos con $\hat{\beta}_1$ y $\hat{\beta}_2$ a los coeficientes estimados, para así distinguirlos de los parámetros "reales" β_1 y β_2, cuyo valor nos gustaría saber. Si logramos una buena estimación de los coeficientes y si el supuesto de media cero de los errores se cumple, entonces podremos hacer un pronóstico acertado.

Figura 1.3: ¿Cuándo se alcanzará 1 US$/Wp?



1.2 Buscando el mejor ajuste: cálculo del estimador MCO

En análisis de regresión contempla un gama de **estimadores** (métodos de estimación) para los coeficientes $\hat{\beta}_1$ y $\hat{\beta}_2$. El más popular de ellos es el método de **mínimos cuadrados ordinarios (MCO)**, también llamado método mínimo cuadrático ordinario. ¿Cómo se estiman $\hat{\beta}_1$ y $\hat{\beta}_2$ con MCO?

El punto de partida es una base de datos con las variables que queremos relacionar. Podemos ordenarla como en la tabla 1.1a. Cada fila i de la tabla se denomina una **observación** y el conjunto de las $n = 13$ observaciones se denomina **muestra**. La producción acumulada en MW y el costo por Wp son las variables que vamos a relacionar.

El estimador MCO fue diseñado para estimar relaciones *lineales*, es decir, rectas (bidimensionales o multidimensionales). Como vemos en la figura 1.1, la relación original entre las variables es no lineal. Si este es el caso usualmente podemos efectuar alguna alteración a los datos o al modelo para lograr la linealidad (más adelante se discute en detalle qué transformaciones a los datos pueden ser útiles para lograr linealidad en casos particulares). En nuestro ejemplo bastó con tomar logaritmos de ambas variables para obtener una relación aproximadamente lineal.

Una vez que ya contamos con las variables que se relacionan linealmente entra en juego el método MCO. Asumimos que entre ellas existe una relación de la forma

$$y_i = \beta_1 + \beta_2 x_i + u_i \qquad \mathrm{E}[u_i] = 0 \ \ \forall i \tag{1.4}$$

y queremos encontrar con los datos de la muestra una función de ajuste,

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i, \tag{1.5}$$

La distinción entre (1.4) y (1.5) es importante. (1.4) representa a la "realidad" que da origen a los datos observados, pero que no es observable en su totalidad, pues sólo

Tabla 1.1: Precio de energía solar fotovoltaica: Ajuste lineal para la curva de aprendizaje

(a) Base cruda				(b) Transformación		(c) Ajuste	
Año	i (# Obs.)	Prod. acum. [MW]	Costo [$/Wp]	x (Log. prod. acum.)	y (Log. costo)	$\hat{y}=2{,}8-0{,}11x$ (Ajuste lineal)	$\hat{u}=y-\hat{y}$ (Residuo)
1998	1	11.3	11.99	2.42	2.48	2.52	-0.04
1999	2	14.2	11.92	2.65	2.48	2.49	-0.01
2000	3	18.1	11.21	2.9	2.42	2.46	-0.04
2001	4	29.4	10.82	3.38	2.38	2.41	-0.03
2002	5	52.1	10.78	3.95	2.38	2.35	0.03
2003	6	97.5	9.79	4.58	2.28	2.28	0
2004	7	155.2	9.16	5.04	2.21	2.23	-0.02
2005	8	234.4	8.79	5.46	2.17	2.19	-0.02
2006	9	339	8.85	5.83	2.18	2.15	0.03
2007	10	508.1	8.95	6.23	2.19	2.1	0.09
2008	11	817.1	8.6	6.71	2.15	2.05	0.1
2009	12	1251.8	8.21	7.13	2.11	2	0.11
2010	13	2164.5	6.97	7.68	1.94	1.94	0
2011	14	3978.2	6.19	8.29	1.82	1.88	-0.06
2012	15	6742	5.31	8.82	1.67	1.82	-0.15

Nota: Precio en dólares del 2012. Fuente: Barbose, Galen, Naïm Darghouth, Samantha Weaver, and Ryan Wiser. 2013. Tracking the Sun VI: An Historical Summary of the Installed Price of Photovoltaics in the United States from 1998 to 2012.

disponemos de x e y en nuestra base de datos, siendo β_1 y β_2 y u desconocidos. En econometría se utilizan dos denominaciones para referirse a esta "realidad": **población** o **proceso generador de datos (PGD)**. En (1.5) lo que se representa es el modelo estimado, cuyos coeficientes estimados $\hat{\beta}_1$, $\hat{\beta}_2$ pueden diferir de sus contrapartes poblacionales β_1 y β_2 (en la figura 1.3 se presentan dos ajustes distintos, basados en métodos de estimación diferentes, pero el PGD que dio origen a los datos sigue siendo el mismo en ambos casos).

Definimos como **residuo** a la diferencia entre cada valor observado de la variable dependiente y su valor pronosticado por la función de ajuste:

$$\hat{u}_i = y_i - \hat{y}_i$$

Así, una representación del modelo que queremos estimar es

$$y_i = \underbrace{\hat{\beta}_1 + \hat{\beta}_2 x_i}_{\hat{y}_i} + \hat{u}_i.$$

La pregunta ahora es: dadas las cantidades que sí podemos observar (x e y), ¿cómo elegimos valores de $\hat{\beta}_1$ y $\hat{\beta}_2$ con el mejor ajuste a los datos observados? Una respuesta natural es buscar valores para los coeficientes que minimicen la distancia entre la recta estimada y la ubicación de los datos en el plano (x,y). Pero existen múltiples formas de minimizar esta distancia, siendo el método MCO una forma particular.

El método de mínimos cuadrados ordinarios lleva su nombre debido a que la función objetivo del problema de optimización a resolver es la suma de los residuos cuadrados $\hat{u}_i^2$. Matemáticamente buscamos:

$$\left\{\hat{\beta}_1, \hat{\beta}_2\right\} = \underset{\hat{\beta}_1,\hat{\beta}_2}{\arg\min} \sum_{i=1}^{n} \hat{u}_i^2 \qquad \text{con} \qquad \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 \tag{1.6}$$

Note que existe una gran diferencia entre "residuo", que es de la estimación, y "error" que es de la población. No podemos minimizar los errores u_i, puesto que son inobservables. Sin embargo, siempre podremos observar el residuo ($\hat{u}_i$): la diferencia entre nuestra recta estimada ($\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$) y el valor observado de y_i.

Si probáramos con distintos valores aleatorios para los coeficientes $\hat{\beta}_1$ y $\hat{\beta}_2$, encontraríamos que algunas combinaciones arrojarían una suma de residuos cuadrados mayores a las de otras combinaciones, tal cual se presenta en la figura 1.4.

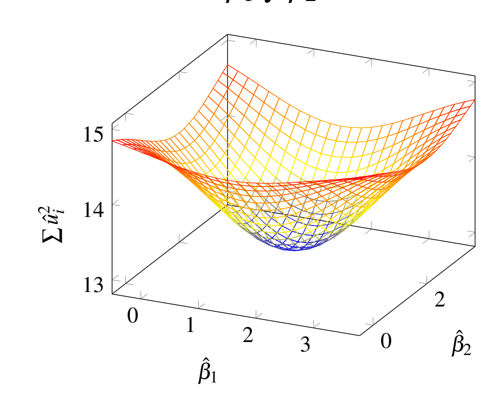
¿Cuáles son los valores exactos de $\hat{\beta}_1$ y $\hat{\beta}_2$ que minimizan la suma de residuos cuadrados? Una forma de obtener la solución sería evaluar

$$\frac{\partial \sum_{i=1}^{n} \hat{u}_i^2}{\partial \hat{\beta}_1} = 0 \qquad \frac{\partial \sum_{i=1}^{n} \hat{u}_i^2}{\partial \hat{\beta}_2} = 0$$

y resolver el sistema de ecuaciones resultante de ambas condiciones de primer orden.

Acá se optará por una solución más general, la que servirá también si se deben estimar más de dos parámetros. Esto puede resultar útil si, retomando el ejemplo anterior, usted cuenta con otros indicadores numéricos que debieran influir en el precio de la energía solar como, por ejemplo, el precio del polisilicio.

Figura 1.4: $\sum \hat{u}_i^2$ para distintas combinaciones de $\hat{\beta}_1$ y $\hat{\beta}_2$



En términos generales, cuando contamos con k variables explicativas (denominados **regresores**) y n observaciones, podemos expresar el modelo de la siguiente manera:

$$y_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2,i} + \ldots + \hat{\beta}_k x_{k,i} + \hat{u}_i \qquad \forall\, i = 1, ..., n \tag{1.7}$$

Como dicha relación se cumple para toda observación i (es decir, para cada año de nuestro ejemplo), también podemos representar (1.7) con vectores:

$$\underset{(n\times1)}{y} = \hat{\beta}_1 \underset{(n\times1)}{1} + \hat{\beta}_2 \underset{(n\times1)}{x_2} + \hat{\beta}_3 \underset{(n\times1)}{x_3} + ... + \hat{\beta}_k \underset{(n\times1)}{x_k} + \underset{(n\times1)}{\hat{u}},$$

cuyos elementos son

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \hat{\beta}_1 + \begin{bmatrix} x_{2,1} \\ x_{2,2} \\ \vdots \\ x_{2,n} \end{bmatrix} \hat{\beta}_2 + \ldots + \begin{bmatrix} x_{k,1} \\ x_{k,2} \\ \vdots \\ x_{k,n} \end{bmatrix} \hat{\beta}_k + \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix}$$

Las variables explicativas pueden ser agrupadas en una sola matriz de dimensión $n \times k$,

la que denotaremos con X.[1] De igual forma, los coeficientes de regresión pueden ser agrupados en un solo vector de dimensión $k \times 1$:

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} 1 & x_{2,1} & x_{3,1} & \cdots & x_{k,1} \\ 1 & x_{2,2} & x_{3,2} & \cdots & x_{k,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2,n} & x_{3,n} & \cdots & x_{k,n} \end{bmatrix}}_{X} \cdot \underbrace{\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}}_{\hat{\beta}} + \underbrace{\begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix}}_{\hat{u}}$$

Así, la forma matricial de representar (1.7) es simplemente

$$\underset{(n\times 1)}{y} = \underset{(n\times k)(k\times 1)}{X \quad \hat{\beta}} + \underset{(n\times 1)}{\hat{u}} \tag{E}$$

Pese a ser un escalar, la sumatoria de residuos cuadrados también tiene una representación matricial. Esta es:

$$\begin{aligned} \sum_{i=1}^{n} \hat{u}_i^2 = \hat{u}'\hat{u} &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &= \left(y' - (X\hat{\beta})'\right)\left(y - X\hat{\beta}\right) && \text{aplicando } (A+B)' = A' + B' \\ &= \left(y' - \hat{\beta}'X'\right)\left(y - X\hat{\beta}\right) && \text{aplicando } (AB)' = B'A' \\ &= y'y - y'X\hat{\beta} - \hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} \\ &= y'y - 2y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \end{aligned}$$

En la última equidad hace uso de $\hat{\beta}'X'y = (\hat{\beta}'X'y)' = y'X\hat{\beta}$, lo que se cumple debido a que un escalar es igual a su transpuesta.

La idea ahora es encontrar un vector $\hat{\beta}$ que contenga los coeficientes que minimicen esta expresión. Es decir, buscamos

$$\hat{\beta} = \arg\min_{\hat{\beta}}[\hat{u}'\hat{u}] = \arg\min_{\hat{\beta}}\left[y'y - 2y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}\right]$$

La minimización se puede obtener derivando $\hat{u}'\hat{u}$ respecto a cada uno de los k coeficientes contenidos en $\hat{\beta}$, igualando a cero y resolviendo el sistema de k ecuaciones con k incógnitas. Por fortuna esta condición de primer orden también se obtiene con ayuda de un par de reglas de diferenciación matricial. Primero notemos que la derivada de $y'y$ respecto a $\hat{\beta}$ debe ser cero, de modo que

$$\frac{\partial \hat{u}'\hat{u}}{\partial \hat{\beta}} = -2\frac{\partial y'X\hat{\beta}}{\partial \hat{\beta}} + \frac{\partial \hat{\beta}'X'X\hat{\beta}}{\partial \hat{\beta}}.$$

[1] Para X utilizaremos notación habitual en econometría: $x_{c,f}$ con c = columna y f = fila. Esta forma discrepa de la notación habitual del álgebra donde los elementos matriciales se suelen representar en el formato $a_{f,c}$.

Luego, si consideramos al vector de $1 \times k$, $a = y'X$, podemos aplicar la regla general $\frac{\partial a_{(1\times k)} z_{(k\times 1)}}{\partial z_{(k\times 1)}} = a'_{(1\times k)}$ (la demostración se remite al apéndice), lo que equivale a $-2X'y$ para el primer sumando.
El segundo término consta de una derivada un poco más compleja. Primero notaremos que el término $X'X$ es una matriz simétrica:

$$
\begin{aligned}
X'X &= \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k,1} & x_{k,2} & \cdots & x_{k,n} \end{bmatrix} \begin{bmatrix} x_{1,1} & x_{2,1} & \cdots & x_{k,1} \\ x_{1,2} & x_{2,2} & \cdots & x_{k,2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,n} & x_{2,n} & \cdots & x_{k,n} \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^n x_{1,i}^2 & \sum_{i=1}^n x_{1,i}x_{2,i} & \cdots & \sum_{i=1}^n x_{1,i}x_{k,i} \\ \sum_{i=1}^n x_{2,i}x_{1,i} & \sum_{i=1}^n x_{2,i}^2 & \cdots & \sum_{i=1}^n x_{2,i}x_{k,i} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{k,i}x_{1,i} & \sum_{i=1}^n x_{k,i}x_{2,i} & \cdots & \sum_{i=1}^n x_{k,i}^2 \end{bmatrix}
\end{aligned}
$$

Siendo $X'X$ simétrica (note que $\sum_{i=1}^n x_{1,i}x_{2,i} = \sum_{i=1}^n x_{2,i}x_{1,i}$, etc.) podemos hacer uso de la regla $\frac{\partial z'Az}{\partial z} = 2Az$, válida siempre y cuando $A_{k\times k}$ sea simétrica (ver demostración en el apéndice).

Con la derivada resuelta, sólo queda despejar el vector $\hat{\beta}$ de

$$
\frac{\partial \hat{u}'\hat{u}}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} = \underset{k\times 1}{0} \tag{1.8}
$$

para obtener el vector de coeficientes estimados resultante del sistema de k ecuaciones que representa (1.8). La solución es:

Fórmula: Estimador mínimo cuadrático ordinario (MCO)

$$
\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = (X'X)^{-1}X'y \tag{1.9}
$$

Un pregunta importante en la práctica es: ¿Bajo qué condiciones existe una solución (1.9) para $\hat{\beta}$?
$X'y$ siempre existe, pero la existencia de $(X'X)^{-1}$ está garantizada sólo si es que se cumple el siguiente supuesto.

Supuesto 1 (S1): La matriz X tiene rango completo.

$$
\text{rango}[X] = k \tag{S1}
$$

Para que (S1) se cumpla, se debe tener (i) $n \geq k$ y (ii) ausencia de **multicolinealidad perfecta** (ninguna columna de X es linealmente dependiente de otra(s) columna(s) de la matriz).

Dicho de otro modo, un único conjunto de parámetros $\hat{\beta}$ es el que minimiza la suma de cuadrados ordinarios siempre y cuando se cumpla S1.

GRETL: ¿Cómo hacer todo esto?

- **Abrir una base de datos**. Para abrir los datos en gretl puede copiar la tabla 1.1a y pegarlos en excel con el siguiente formato (nombres en primera fila, sin espacios ni caracteres conflictivos):

	A	B	C	D
1	ano	i	ProdAcum	Costo
2	1998	1	11.3	11.99
3	1999	2	14.2	11.92
4	2000	3	18.1	11.21

  Grabe la base de datos, por ejemplo, como `C:\SOLAR.xlsx`. Abra gretl y pinche el segundo ícono de abajo a la izquierda para abrir un 'guión nuevo'. En él escriba el siguiente comando:

```
open "C:\SOLAR.xlsx"
```

  Para ejecutarlo presione Ctrl+r.
- **Graficar variables**. Pueden utilizarse los comandos `gnuplot` o `scatters`:

```
gnuplot Costo ProdAcum --output=display --suppress-fitted
scatters Costo; ProdAcum --output=display
```

- **Crear una variable**. El comando `series` sirve para crear una nueva variable. Por ejemplo, los logaritmos se crean con:

```
series l_Costo = ln(Costo)
series l_ProdAcum = ln(ProdAcum)
```

- **Regresión MCO**. Por su siglas en inglés (*ordinary least squares*), el comando en gretl es `ols`. Por ejemplo:

```
ols l_Costo const l_ProdAcum
```

  Note que `const` representa a la constante (un vector con unos).
- **Ajuste y residuos**. Tras efectuar el comando `ols` es posible acceder al ajuste y a los residuos mediante:

```
series Ajuste = $yhat
series Residuos = $uhat
```

Incluya todos estos comandos a su guión y ejecútelo.

1.3 ¿Cómo afecta la inequidad el desarrollo económico de los países?

En estos días ha surgido un debate en torno a los efectos que tiene la inequidad sobre el desarrollo económico de las naciones. La visión tradicional, basada en modelos teóricos de cómo afectan los impuestos al crecimiento económico, es que la inequidad no debe ser combatida con redistribución. Algunos economistas, sin embargo, arguyen que la relación es más compleja, especialmente en el largo plazo, y que no hay que sacar conclusiones antes de verificar qué dicen los datos al respecto.

¿Cómo podríamos verificar la relación largoplacista que existe entre inequidad y desarrollo económico?

Una alternativa simple sería revisar cómo se relaciona el nivel de inequidad en, por ejemplo, 1950 con el ingreso per cápita actual. En la tabla 1.2 se presentan datos de los pocos países para los cuales se tiene dicha información y en la figura 1.5 se presenta gráficamente la relación entre ambas variables.

Al parecer, podríamos postular que cada país $i = 1,...,13$ obedece a una relación lineal entre ambas variables de la forma

Tabla 1.2: Base de datos con dos regresores

	i (# Obs.)	x_2 (Inequidad en 1950)	x_3 (Ingreso en 1950)	y (Ingreso actual)
Argentina	1	2.50	4934.4	9527.6
Australia	2	1.80	7276.4	24805.4
Canadá	3	1.80	7438.6	24886.0
Dinamarca	4	1.99	6581.6	24233.8
Francia	5	1.80	4901.4	21712.1
Alemania	6	2.23	3988.3	20209.6
India	7	2.79	638.2	2887.2
Japón	8	1.71	2006.0	21832.7
Malasia	9	2.19	1405.1	9527.3
Noruega	10	1.83	5361.5	28030.0
Singapur	11	2.17	2290.5	26189.6
Suecia	12	1.72	6563.0	24661.9
Suiza	13	2.09	9354.4	24605.5

Nota: Las variables son promedios de 10 años en torno a la fecha indicada. El índice de inequidad es el coeficiente inverso de Pareto-Lorenz de la Top Income Database. El PIB real per cápita es del Maddison Proyect.

$$\underbrace{\text{Nivel actual de ingreso}}_{y} = \beta_1 + \beta_2 \underbrace{\text{Inequidad en 1950}}_{x} + \underbrace{\text{Otros factores}}_{u}, \qquad (1.10)$$

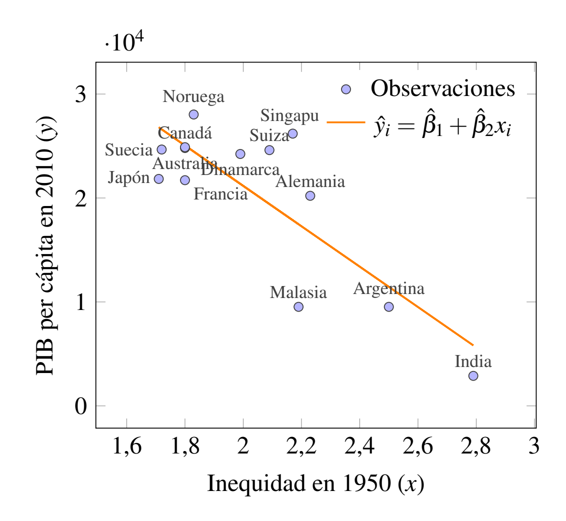
donde el índice de inequidad es la variable explicativa o regresor (x), y el nivel de ingreso per cápita, es la variable dependiente (y). Aquello que se escapa a la relación lineal entre x e y se captura en el error (u). Dentro del error podrían estar, por ejemplo, políticas de estado, acceso a recursos naturales, así como cualquier variable que afecte al nivel de ingreso per cápita.

Note que en (1.10) se surgiere implícitamente una **causalidad**: y es función de x. Este es un supuesto importante que discutiremos más adelante.

Si asumimos que los "otros factores" son cero en la media y que se cumple la premisa de causalidad de x hacia y, entonces podemos plantear

$$\mathrm{E}[y_i] = \beta_1 + \beta_2 x_i. \qquad (1.11)$$

Figura 1.5: Regresión lineal



Es decir, con conocimiento de β_1 y β_2, podríamos predecir el nivel de ingreso más probable que tendría un país del cual sabemos cuál fue su nivel de inequidad en 1950. Si β_2 es en realidad un parámetro negativo, entonces aquellos países que combatieron la inequidad en el pasado tienen hoy un nivel de ingreso per cápita superior. Esto podría servir de ejemplo para adoptar políticas que reduzcan la inequidad en el presente para aumentar el nivel de ingreso de un país en el futuro. En cambio, si β_2 es positivo, entonces la redistribución generaría un deterioro del ingreso per cápita del país, tal como indica la teoría económica más convencional. Por ello, más allá de discusiones éticas, conocer el valor de β_2 es de gran importancia para poder elegir la política óptima de crecimiento económico de un país.

Sigamos con la pregunta ¿cuál es el efecto de la inequidad sobre el nivel de ingreso en el largo plazo? e intentemos resolverla con los datos de la tabla 1.2. El primer planteamiento para llegar a una respuesta fue correr la regresión



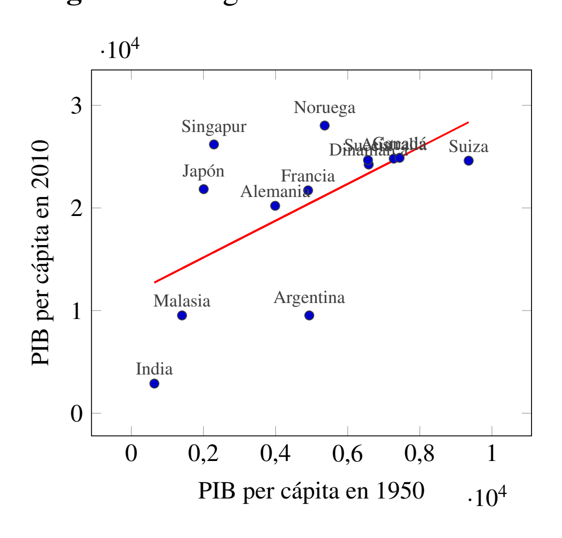
$$\text{PIB}_{2010} = \hat{\beta}_1 + \hat{\beta}_2 \text{Inequidad}_{1950} + \hat{u}, \tag{M1}$$

cuyo resultado, siendo $X = \begin{bmatrix} 1 & x_2 \end{bmatrix}$, es

$$\hat{\beta} = (X'X)^{-1}X'y = \begin{bmatrix} 5{,}99e+04 \\ \text{-}1{,}94e+04 \end{bmatrix}$$

El resultado $\hat{\beta}_2 < 0$ no nos sorprende dado que ya vimos esta relación gráficamente en la figura 1.5.

Figura 1.6: Ingreso en 2010 vs. 1950



Pero, ¿no será que la correlación entre el ingreso actual y la inequidad en 1950 se debe simplemente a que países con mayor nivel de ingreso tienen menor inequidad, y los países con poca inequidad ya tenían un ingreso alto en 1950? (Si no entendió esto, vuelva a leerlo)

En la figura 1.6 se presenta la relación entre en ingreso actual y el ingreso pasado de los países de la muestra junto con la estimación de

$$\text{PIB}_{2010} = \hat{\beta}_1 + \hat{\beta}_2 \text{PIB}_{1950} + \hat{u}. \tag{M2}$$

En efecto vemos que la **correlación** es positiva ($\hat{\beta}_2 > 0$), es decir, países de alto ingreso de hoy ya tenían un alto ingreso en 1950. Por ende, resulta razonable pensar que la correlación entre el PIB de 2010 y la inequidad de 1950 es **espuria**: se debe a que ambas variables se correlacionan con una tercera variable explicativa, el nivel de ingreso en 1950.

¿Cómo saber, entonces, si la correlación entre la inequidad pasada y el ingreso actual es "independiente" de el nivel de ingreso pasado. La forma de dar respuesta a la interrogante es correr una regresión con las dos variables explicativas. El modelo a estimar sería

$$\text{PIB}_{2010} = \hat{\beta}_1 + \hat{\beta}_2 \text{Inequidad}_{1950} + \hat{\beta}_3 \text{PIB}_{1950} + \hat{u}. \qquad \text{(M3)}$$

Esto es lo que se denomina "controlar" por la variable PIB_{1950} (en este contexto PIB_{1950} es un **control**). Más adelante veremos cómo incluir controles de primer orden es fundamental para un buen análisis econométrico.

Como en M3 $X = \begin{bmatrix} 1 & x_2 & x_3 \end{bmatrix}$, el estimador para $\hat{\beta}$ dará un resultado distinto. En términos generales, si el resultado para un coeficiente de interés se mantuviera relativamente constante pese a la inclusión de controles, cambios en el número de observaciones, cambios en el periodo analizado, etc., se habla de un resultado **robusto** (pues no depende de la método particular de estimación).

Tabla 1.3: Regresiones MCO en base a la tabla 1.2

Variable dependiente: Ingreso en 2010

Modelo	(M1)	(M2)	(M3)
const	5.99e+04	1.16e+04	4.92e+04
Inequidad en 1950	−1.94e+04		−1.62e+04
Ingreso en 1950		1.79	0.848
n	13	13	13
R^2	0.636	0.314	0.731

Los resultados de las tres regresiones M1, M2 y M3 se presentan en la tabla 1.3.

Una primera conclusión es que el coeficiente de Inequidad_{1950} mantuvo su signo y el orden de magnitud, es decir, resultó ser relativamente robusto al control. El coeficiente de PIB_{1950}, en cambio, se redujo en más de 50 %. Sin embargo, se mantuvo positivo.

¿Cómo interpretamos los resultados de M3? En la figura 1.6 vemos países que están sobre la recta y bajo la recta estimada. Argentina, por ejemplo, parece haber crecido poco dado su nivel de ingreso, al igual que la India y Malasia. Singapur y Japón, en tanto, se encuentran por sobre la recta, indicando que crecieron más de lo esperado según el modelo estimado M2. Volviendo a la figura 1.5 vemos que Argentina y la India tenían altos niveles de Inequidad_{1950} mientras Japón tenía niveles bajos. Es decir, para algunos

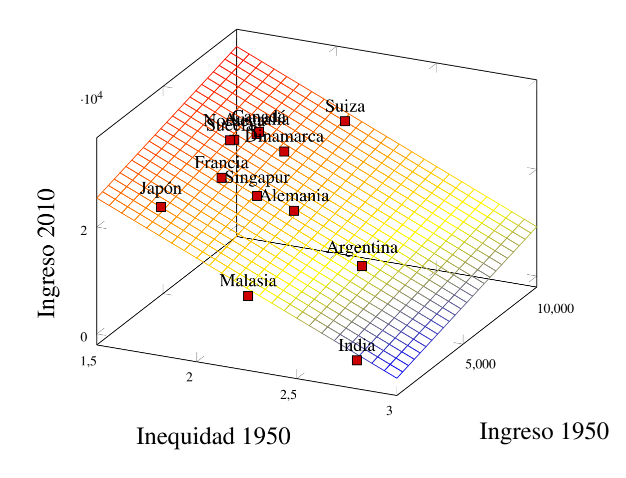
países, lo que no se pudo "explicar" con PIB_{1950} sí se puede explicar con Inequidad_{1950}. De la misma manera, como se aprecia en la figura 1.5, Alemania está por sobre la recta del modelo M1 y Malasia está por debajo. Pero esa diferencia se puede explicar con PIB_{1950}: Alemania tenía mayor ingreso que Malasia.

Así, el ajuste de M3,

$$\widehat{\text{PIB}}_{2010} = 4{,}92e+04 - 1{,}62e+04\text{Inequidad}_{1950} + 0{,}848\text{PIB}_{1950},$$

nos entrega una relación multidimensional entre la variable dependiente y los regresores, tal cual se presenta en la figura 1.7. Malasia, por ejemplo, pese a tener un nivel de ingreso similar al de la India, logró un mayor ingreso en 2010, el que se explica por una menor inequidad. Argentina, pese a tener un nivel de ingreso comparable al de Francia o Alemania, no se siguió desarrollando debido a los niveles de inequidad. Singapur, pese a ser más equitativo que Suiza, no ha logrado el nivel de ingreso de los Suizos debido a que su ingreso medio en 1950 era muy bajo.

Figura 1.7: Representación de M3



Con esta lógica, los coeficientes $\hat{\beta}_j$ son denominados **coeficientes de correlación parcial**. Su interpretación es: cuánto estimamos que cambia y si aumentamos x_j en una unidad y el resto de los regresores se mantienen constantes. Es decir, proveen un estimador de

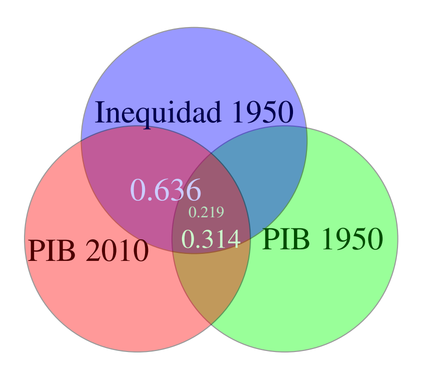
$$\beta_j = \frac{\partial y}{\partial x_j}.$$

Note que en M3 tanto el coeficiente de Inequidad_{1950} como el de Ingreso_{1950} vieron reducido su impacto respecto a los modelo M1 y M2 (es decir, el impacto se acerca a cero en ambos casos). ¿Por qué? El resultado es común en la econometría y se da cuando existe cierto nivel de correlación **entre** las variables explicativas (esto se denomina **multicolinalidad imperfecta**). La idea se puede representar gráficamente con ayuda del diagrama de Venn. Los círculos de la figura (1.8) representan la variación total de cada una de las variables. Las intersecciones corresponden a la varianza común que pueda existir entre ellas. Como parte de la variación común entre Ingreso_{2010} e Inequidad_{1950} es al mismo tiempo una variación común con Ingreso_{1950}, el impacto que se atribuye a cada uno de los coeficientes en M3 es menor.

Así, si incluyeramos una tercera variable que se encontrara correlacionada con la variable dependiente y con las variables explicativas, los coeficientes cambiarían nuevamente. ¿Perderán impacto Inequidad_{1950} e Ingreso_{1950}? Aunque es un escenario probable, la respuesta no es clara, pues depende de condiciones que estudiaremos más adelante. Lo que sí sabemos es que la estimación cambiará.

Las conclusiones expuestas acá naturalmente son fuente de controversia y no deben ser tomadas como una verdad irrefutable, sino como un ejemplo de la aplicación del método MCO, su alcance y sus limitantes. Como veremos más adelante, todo resultado econométrico es altamente dependiente de cuán correcta sea la especificación del modelo estimado y, naturalmente, de la representatividad de la muestra que se utiliza. ¿Cómo cambian los resultados si se incluyeran observaciones para países como Polonia o Yugoslavia? ¿Cambian los resultados al tomar otros años de referencia para los regresores o la variable dependiente? ¿Se mantiene el resultado con medidas alternativas de inequidad? Un trabajo serio de investigación debe intentar dar respuesta a este tipo de interrogantes en un **análisis de robustez**.

Figura 1.8: Diagrama de Venn



GRETL: Presentación de múltiples modelos

Para presentar múltiples regresiones existe el comando `modeltab`:

```
modeltab free # para limpiar la tabla
ols ingr2010 const ineq1950
modeltab add
ols ingr2010 const ingr1950
modeltab add
ols ingr2010 const ineq1950 ingr1950
modeltab add
modeltab show # para mostrarla
```

1.4 Bondad de ajuste y causalidad

Para cada modelo de la tabla 1.3 se reporta en la última fila el R^2 o **coeficiente de determinación**. El R^2 corresponde a una medida de **bondad de ajuste** que responde a la pregunta: ¿qué fracción de la dispersión total de y es explicada por la recta (multidimensional) estimada $X\hat{\beta}$?

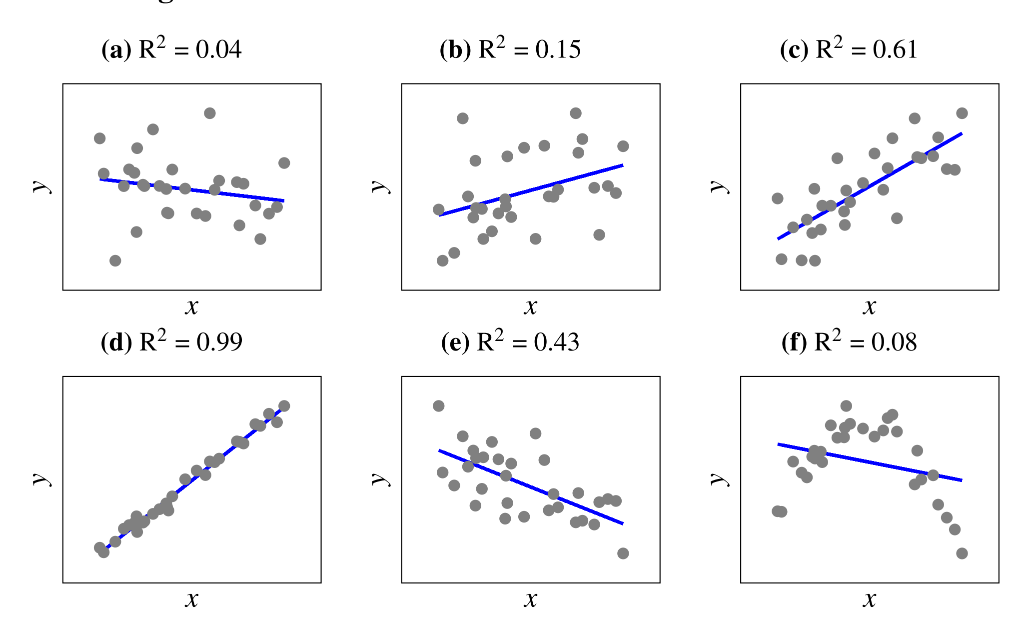
De todas las representaciones del R^2, la más intuitiva es:

$$
\begin{aligned}
R^2 &= 1 - \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= 1 - \frac{\sum_{i=1}^n (\hat{u}_i - \bar{\hat{u}})^2/n}{\sum_{i=1}^n (y_i - \bar{y})^2/n} = 1 - \frac{\mathrm{var}[\hat{u}]}{\mathrm{var}[y]} \qquad (\text{note que } \bar{\hat{u}} = 0)
\end{aligned}
$$

Si el modelo explica poco, $\sum_{i=1}^n \hat{u}_i^2$ es alto y el R^2 es bajo. En el peor de los casos la suma de residuos cuadrados son equivalente a la dispersión de la variable dependiente, que está dada por $\sum_{i=1}^n (y_i - \bar{y})^2$, y el R^2 es cero. En el mejor de los casos todos los residuos son cero y el R^2 es uno. La figura 1.9 muestra el valor que toma el coeficiente de una recta ajustada $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$ en distintas situaciones. Mientras en 1.9d un 99 %

de la varianza de y se captura con el modelo, en 1.9a sólo un 4 % de dicha varianza se explica por el modelo y un 96 % queda inexplicada en forma de residuos.

Figura 1.9: Coeficiente de determinación en distintos casos



En el ejemplo de la tabla 1.3, como M3 explica parte de los residuos (es decir lo no explicado) de M1 y M3, el R^2 más alto de la tabla es naturalmente el de M3. Como regla general, tras introducir un regresor adicional a una regresión, el R^2 siempre será igual o mayor. Sin embargo, note que el R^2 del modelo M3 no corresponde a la suma de los R^2 de los modelos M1 y M3. ¿Por qué? La respuesta se encuentra en el diagrama de Venn (figura 1.8): 0.219 de lo que se explica con M1 también se explica con M2, llegando el nuevo R^2 a $0{,}636+0{,}314-0{,}219=0{,}731$.

Podemos decir que en M3 se explica el 73 % de la dispersión del ingreso medio de los países tan sólo con dos regresores: el ingreso medio del país en 1950 y su desigualdad de ingresos en ese mismo año. En otras palabras, al incluir otras variables explicativas como, por ejemplo, la abundancia de recursos productivos, sólo podríamos mejorar en un máximo de 27 % nuetra explicación de y.

Pero hay que ser cuidadosos, porque la capacidad de "explicar" algo que sucedió no es equivalente a la capacidad de predecir qué va a ocurrir en el futuro. Por motivos que quedarán claros más adelante, incluso un modelo con un R^2 de 1 puede tener nula capacidad predictiva (es decir, el modelo sería inservible). Mientras que otro modelo con un R^2 de tan sólo 30 % puede tener una exelente capacidad predictiva.

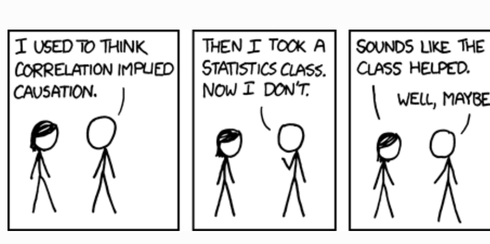
Para ilustrar cómo la bondad de ajuste o "variación explicada" puede ser un engaño, considere la posibilidad de que los países que han logrado un elevado nivel de ingreso per cápita lo han logrado gracias a un espíritu colectivista que mejora la productividad de las empresas por medio de un buen trabajo en equipo. Suponga además que quienes comparten ese espíritu también votan a favor de políticas redistributivas. ¿Cuál va a ser

el resultado? La baja inequidad aparecería como buena variable explicativa debido a que está operando como **proxy** de la variable fundamental, que sería colectivismo. Si un país aumenta la equidad pero el colectivismo se mantiene inalterado, entonces la política redistributiva no tiene ningún efecto sobre el nivel de ingreso. En conclusión, "la bondad de ajuste" no sería ningún indicador de la "bondad del modelo", debido a la presencia de una correlación espuria. Un alto R^2 es sinónimo de una alta correlación, pero correlación no siempre implica causalidad.

Origen de una correlación

1. Casualidad directa $\longrightarrow$ **x** causa **y**.
2. Causalidad inversa $\longrightarrow$ **y** causa **x**.
3. Causalidad simultánea $\longrightarrow$ **x** causa **y** e **y** causa **x**.
4. Correlación espuria $\longrightarrow$ Tanto **x** como **y** son causa de un factor común **z**.
5. Correlación casual $\longrightarrow$ No hay causalidad.

Figura 1.10: Pensamiento crítico



Sacado de xkcd.com

Para ilustrar la relación entre causalidad, correlación y proxy, la figura 1.11 muestra el ajuste

$$\widehat{\text{precio}} = 877{,}83 - 0{,}433\text{año} \qquad (1.12)$$
$$(R^2 = 0{,}95)$$

Según esta estimación el precio de 1 dólar por watt se alcanzaría en torno al año 2028. El R^2 de la regresión es superior al de la regresión $\ln(\widehat{\text{precio}}) = 2{,}8 - 0{,}13\ln(\text{prod. acum})$ de la de la figura 1.3, cuyo valor es $R^2 = 0{,}92$.

¿Qué modelo otorga una predicción más confiable?

Figura 1.11: Precio del watt fotovoltaico

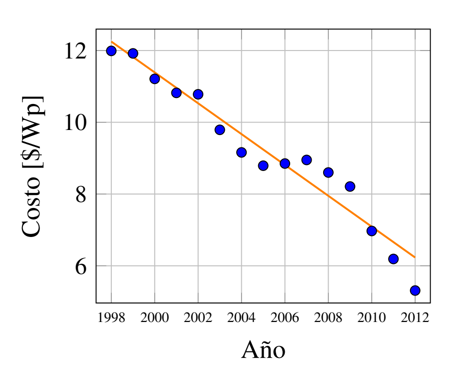
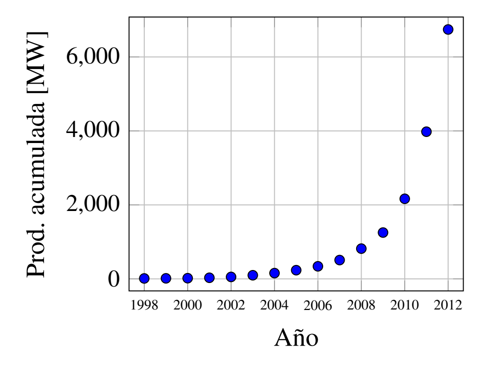


Figura 1.12: Producción fotovoltaica

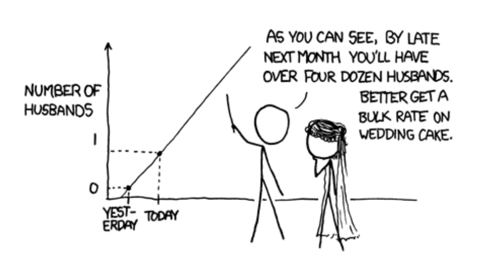


En la figura 1.12 se muestra cómo ha crecido la producción acumulada de watts producidos con energía solar fotovoltaica a lo largo de los años. Como vemos, la correlación es positiva. ¿Qué pasaría con el costo de la tecnología si se dejara de producir energía solar durante los próximos 10 años? Es de esperar que se frene la curva de aprendizaje, no se acumule conocimiento, y el precio se mantenga pese al transcurso del tiempo. Como "año" es una variable proxy de la producción, no

existe certeza de que en el año 2028 tengamos precios competitivos si es que no se sigue invirtiendo esfuerzos en mejorar la tecnología. En (1.4) se estimó una correlación espuria.

De forma irónica, en la figura 1.13 se presenta la proyección del número de matrimonios que tendrá la novia si siguiera con uno al día. El ejemplo es obviamente una exageración, y la novia no tendrá que comprar tantas tortas de matrimonio, pero la idea se aplica a problemas prácticos que a veces no son tan obvios. Incluso especialistas en ocasiones caen en la trampa de interpretar correlaciones como causalidad.

Figura 1.13: Proyección causal



Sacado de xkcd.com

En conclusión, el R^2 sirve para evaluar la bondad de ajuste (corresponde a un indicador de cuán importantes son las variables omitidas para la determinación de y, pues dicha importancia repercute en la varianza del error y en la de los residuos, como también podría a ayudar a detectar problemas como, por ejemplo, el del gráfico 1.9f, donde se estima una relación lineal cuando no corresponde) pero no es un criterio robusto para elegir el mejor modelo, especialmente cuando el objetivo es hacer una predicción fuera del rango de los valores X observados.

En fin, siempre recuerde: una correlación alta (un alto R^2) no implica de causalidad.

GRETL: Coeficiente de determinación

El comando asociado es `$rsq`. Puede utilizarlo de dos maneras. La primera es tras estimar una regresión MCO, por ejemplo:

```
ols y const x1 x2 x3
scalar Rcuadrado = $rsq
```

La segunda es dando nombre a la regresión:

```
MiReg <- ols y const x1 x2 x3
scalar Rcuadrado = MiReg.$rsq
```

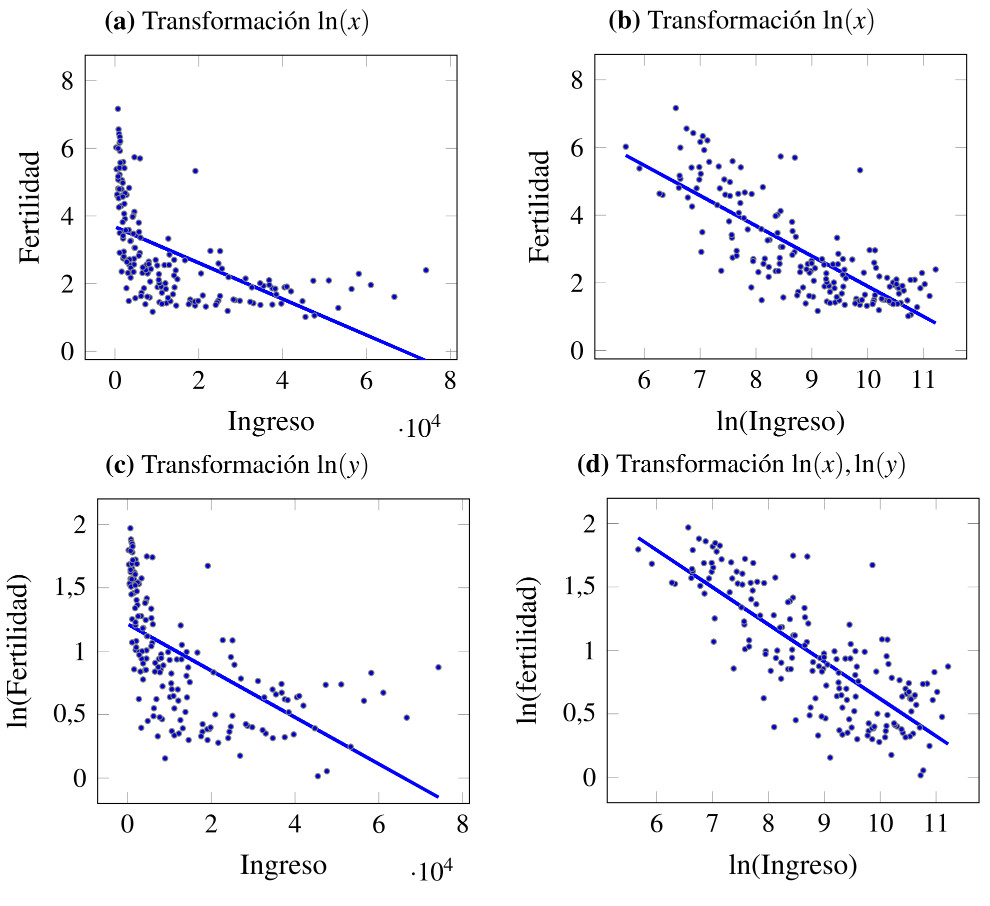
1.5 Transformaciones para la linealidad

Muchas veces deseamos estudiar la relación entre dos o más variables cuya relación no es lineal. Un ejemplo es el de la figura 1.14a. Uno podría estar interesado en predecir la fertilidad de un país en vías de desarrollo considerando que en algunas décadas más su nivel de ingreso per cápita será superior y la tenencia de hijos probablemente se asimilaría más a la de un país desarrollado. Si ya se tiene una predicción del ingreso medio que tendrá el país, la pregunta es qué parámetros gobiernan la relación

$$\text{Fertilidad} = f(\text{Ingreso}).$$

Claramente, fertilidad e ingreso no tienen una relación lineal. ¿Qué se puede hacer para aplicar el modelo MCO en este caso?

Figura 1.14: Transformaciones logarítmicas en x e y



Fuente: Banco Mundial

Una posible solución consiste en transformar las variables Fertilidad e Ingreso de manera tal que se obtenga una relación lineal. Es decir, buscamos funciones $g(\cdot)$ para ambas variables tales que

$$g_y(\text{Fertilidad}) = \beta_0 + \beta_1 g_x(\text{Ingreso}) + u.$$

¿Qué forma funcional elegir para $g_y(\cdot)$ y $g_x(\cdot)$? Para linealizar la relación de la figura 1.14a aplicaremos en ambos ejes un función que cumpla con $g'(\cdot) > 0$ y $g''(\cdot) < 0$. Funciones que cumple con esta característica son, por ejemplo, $g(z) = \ln(z)$, $g(z) = \sqrt{z}$ y $g(z) = z^{\frac{3}{4}}$. Todas ellas tienen en común que acercan en mayor proporción a los puntos alejados del eje y en menor proporción a los puntos cercanos al eje, tal cual se representa en la figura 1.15.

Lo habitual es elegir a $g(\cdot) = \ln(\cdot)$ en lugar de otra función similar. Esto se debe a que una regresión de la formas **log-log**, **nivel-log** y **log-nivel** tendrá una de las prácticas interpretaciones presentadas en la tabla 1.4.

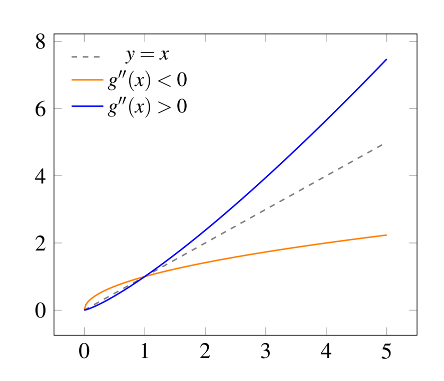
La figura 1.14 muestra cómo con cambia el ordenamiento de los datos al aplicar $g = \ln(\cdot)$, primero en el eje vertical , luego en el eje horizontal y, por último, en ambos

Tabla 1.4: Transformaciones comunes

Denominación	Especificación	Diferencial total	Interpretación
nivel-nivel	$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$	$\Delta\hat{y} = \hat{\beta}_2 \Delta x$	Si x aumenta en una unidad, entonces $\hat{y}$ aumenta en $\hat{\beta}_2$ unidades.
log-log	$\ln \hat{y} = \hat{\beta}_1 + \hat{\beta}_2 \ln x$	$\frac{\Delta\hat{y}}{\hat{y}} = \hat{\beta}_2 \frac{\Delta x}{x}$	Si x aumenta en 1 %, entonces $\hat{y}$ aumenta en $\hat{\beta}_2$ %
nivel-log	$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 \ln x$	$\Delta\hat{y} = \hat{\beta}_2 \frac{\Delta x}{x}$	Si x aumenta en 1 %, entonces $\hat{y}$ aumenta en $\hat{\beta}_2/100$ unidades.
log-nivel	$\ln \hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$	$\frac{\Delta\hat{y}}{\hat{y}} = \hat{\beta}_2 \Delta x$	Si x aumenta en una unidad, entonces $\hat{y}$ aumenta en $100\hat{\beta}_2$ %

ejes a la vez.

Figura 1.15: Transformación $g(\cdot)$ con $g'(\cdot) > 0$ y $g''(\cdot) < 0$



La línea de regresión de la figura 1.14d es

$$\ln(\widehat{\text{Fertilidad}}) = 3{,}55 - 0{,}29 \ln(\text{Ingreso}).$$

Los resultados se interpretan así: un aumento en, por ejemplo, 10 % del ingreso de un país genera una disminución de la tasa de fertilidad en 2,9 %; duplicar el ingreso, por ejemplo, hace caer la tasa de fertilidad a, aproximadamente, 2/3 de su nivel inicial.

Por supuesto, la transformación logarítmica no siempre es adecuada. Un ejemplo es el que se presenta en la figura 1.16a. Si bien aplicar la transformación logarítmica en el eje x ayuda a linealizar la relación, el eje y necesita la transformación inversa, es decir un operador que cumpla con $g''(\cdot) > 0$. En la figura 1.16b se muestra el ajuste lineal aplicando $g(y) = y^2$.

GRETL: Transformaciones

Vía el comando `series` es posible utilizar las funciones `ln(x), exp(x), x^2, x^0.5`, etc. Además existen transformaciones rápidas para múltiples series a la vez. Por ejemplo

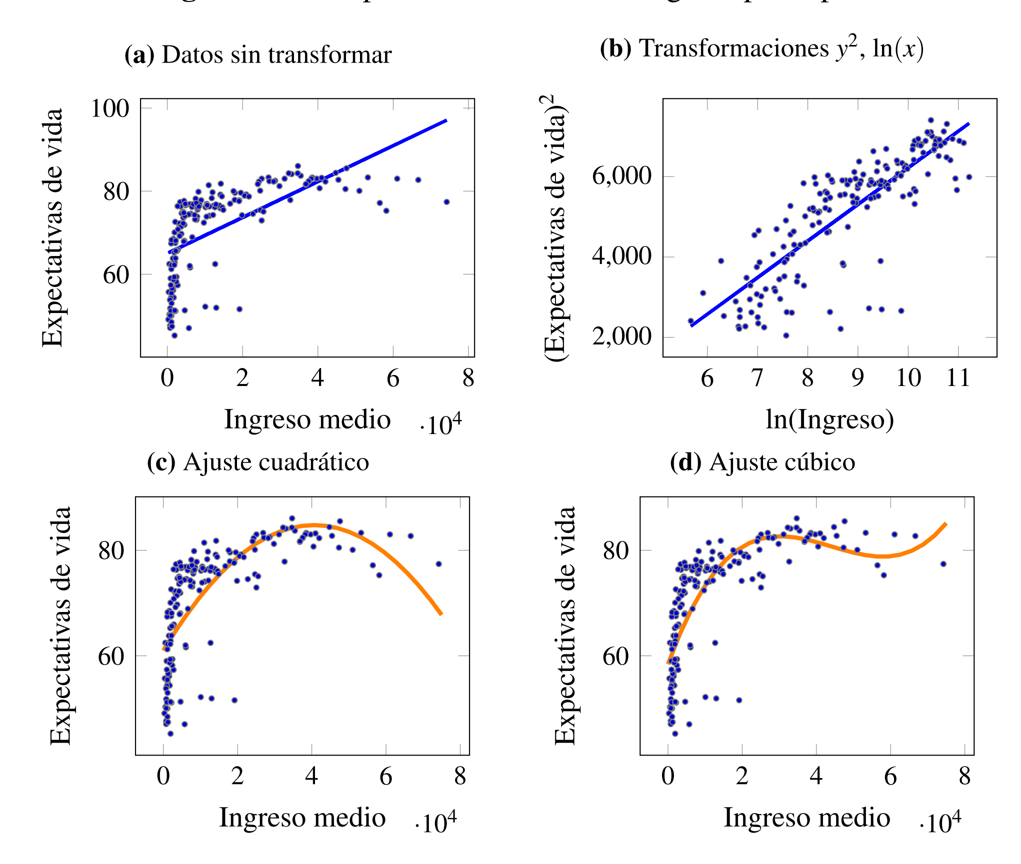
```
logs Costo ProdAcum
squares  Costo ProdAcum
```

crea los logaritmos de ambas variables y sus cuadrados (con prefijos `l_` y `sq_` respectivamente).

Otra alternativa cuando las relaciones no son lineales es estimar una **regresión polinomial** de orden m:

$$y_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{\beta}_3 x_i^2 + \ldots + \hat{\beta}_k x_i^{k-1} + \hat{u},$$

Figura 1.16: Expectativas de vida vs. ingreso per cápita



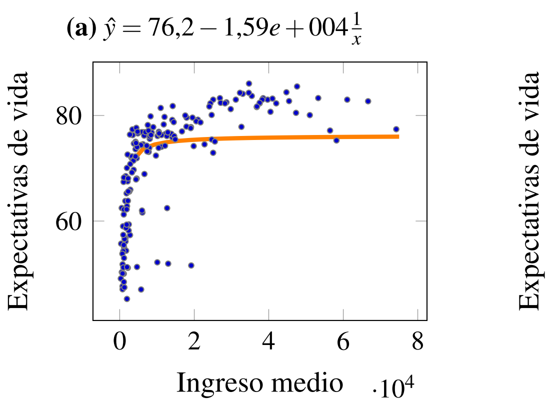
donde $m = k - 1$ en este caso. Las figuras 1.16c y 1.16d presentan ajustes polinomiales de orden 2 y 3 respectivamente. A medida que aumentamos el orden m de la regresión polinomial siempre mejora la calidad del ajuste intra muestra.

Otro tipo de ajuste común es el ajuste inverso:

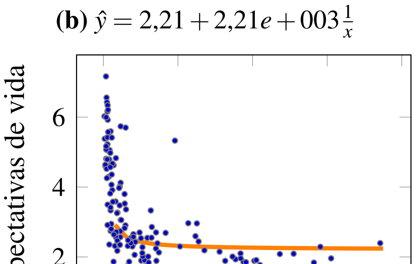
$$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 \frac{1}{x} \tag{1.13}$$

Note que en (1.13) el valor asintótico de la variable dependiente es $\hat{y} \underset{x \to \infty}{\longrightarrow} \hat{\beta}_1$.
En la figura 1.17 se presenta el ajuste para dos de nuestros ejemplos. Claramente el ajuste no es de la mejor calidad para los países con alto nivel de ingreso. Este es un problema habitual del ajuste inverso.

Figura 1.17: Ajuste inverso: $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 \frac{1}{x}$



(a) $\hat{y} = 76{,}2 - 1{,}59e + 004\frac{1}{x}$



(b) $\hat{y} = 2{,}21 + 2{,}21e + 003\frac{1}{x}$

1.6 ¡Esos outliers!

Vimos que un alto R^2 no necesariamente implica un buen modelo. En particular, si la correlación es espuria, no hay causalidad directa y la predicción pierde validez. En muchos casos es difícil establecer si la causalidad va efectivamente de X a y, pero en otros casos es fácil. Sabemos que la nubosidad es un predictor de la lluvia, como sabemos que el metraje de una propiedad determina su precio y no viceversa. Restringiéndonos a casos en que la causalidad está clara, ¿es el R^2 un buen indicador de la bondad del modelo?

La respuesta se encuentra en la figura 1.18, donde presenta 4 regresiones MCO con el mismo R^2. Este cuarteto, bautizado en honor a Francis Anscombe, (i) muestra cómo el R^2 puede ser engañoso para juzgar la calidad del ajuste de un modelo y (ii) ilustra algunas debilidades de la regresión MCO .

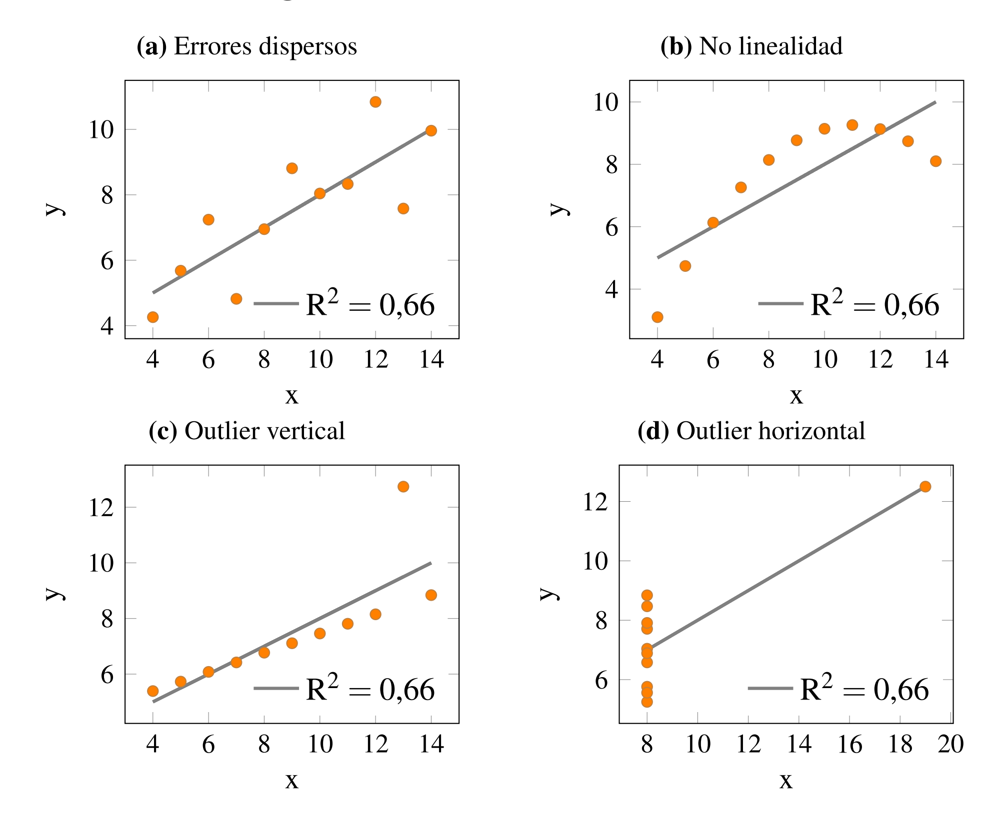
¿Por qué es el R^2 en 1.18a igual que en los otros casos si la relación parece bien capturada por la recta? La razón se encuentra en que los errores tienen una alta varianza. Esto suele se el caso cuando se dejan muchas variables dentro del término de error. Aunque el modelo estimado acá sea correcto, el R^2 es bajo.

El problema de 1.18b es la no linealidad de la relación. Por ahora queda claro que graficar los datos es importante para no cometer este tipo de error en la práctica. Más adelante veremos una metodología para detectar el problema cuando existen muchos regresores y la detección gráfica se dificulta.

En las figuras 1.18c y 1.18d se presentan **outliers**. Ese es el nombre que se le da a observaciones que distan de la relación típica observada entre los datos. Note que se hace una diferenciación entre outliers verticales y horizontales, pues el efecto que tienen sobre una estimación es desigual.

La peor regresión de la figura 1.18 es sin duda la última. Esta regresión padece de un problema denominado **valor influyente**. Llamaremos valor influyente a un punto que cuenta con dos características:

Figura 1.18: El cuarteto de Anscombe



1. Es un outlier: se encuentra considerablemente alejado de la recta poblacional.
2. Tiene alto **apalancamiento**: se encuentra considerablemente alejado de la media en el dominio del eje x.

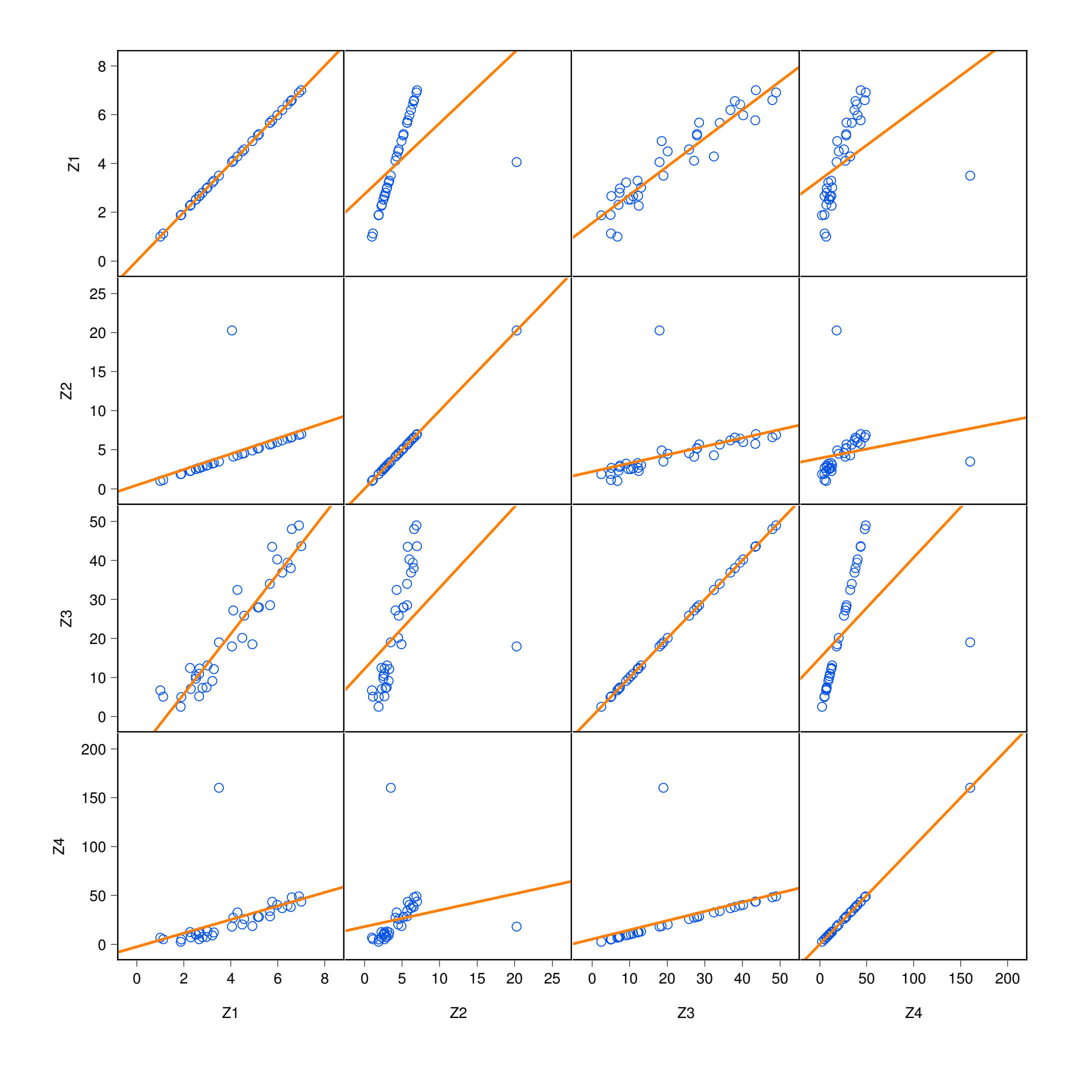
Para ilustrar como son sólo los valores influyentes los que inciden fuertemente en una estimación, la figura 1.19 presenta una serie de estimaciones MCO, donde cada recta se obtuvo mediante una regresión con la variable del eje vertical como variable dependiente y la variable del eje horizontal como regresor.
Vemos que no en todos los casos se distorsiona la estimación, pues se tienen que cumplir 1 y 2 para ello. Por ejemplo, la regresión de Z3 respecto a Z2 se ve fuertemente distorsionada por el outlier mientras la regresión de Z3 respecto a Z3 no presenta alteración en la pendiente, pese a que los datos son exactamente los mismos, sólo que cambiados de la mano izquierda de la regresión a la mano derecha y viceversa. El mismo fenómeno se observa también entre Z4 y Z3, donde se ve que cuando Z4 es variable dependiente el único efecto del outlier es un pequeño cambio en la constante (la recta está levemente más arriba).

Cuando se realizan regresiones entre dos variables es fácil verificar la presencia de puntos de apalancamiento gráficamente. ¿Cómo se hace cuando $k > 2$, es decir, cuando hay más de un regresor?

El apalancamiento se puede cuantificar con ayuda de la ***hat matrix***,

$$\underset{n\times n}{H} = X(X'X)^{-1}X', \tag{1.14}$$

Figura 1.19: Outliers verticales vs. outliers horizontales



cuyos n elementos diagonales $h_i = H_{ii}$ se caracterizan por:

- Su valor está entre cero y uno
- Su valor aumenta con la distancia de X respecto de la media de X.

Las figuras 1.21 y 1.20 ilustran la relación que existe entre el apalancamiento y la distancia de la media de X. Japón tiene un alto nivel de h_i debido que tiene el menor nivel de inequidad en 1950, Suiza tiene un alto nivel por tener el mayor nivel de ingreso en 1950 y la India tiene el mayor nivel de h_i debido a que tiene tanto el mayor nivel de inequidad en 1950 como el menor ingreso en 1950.

Una pregunta natural es: ¿qué niveles de h_i se consideran preocupantes? Algunos autores sugieren revisar qué ocurre con datos que tienen $h_i > 2 \cdot k/n$. En el caso de la regresión M3, $2 \cdot k/n = 2 \cdot 3/12 = 0{,}5$. El único país con $h_i > 0{,}5$ es la India. Esto significa que sería recomendable intentar incorporar un nuevo país a la base de datos con características similares a la India para ver si los resultados de M3 son robustos.

Pero, tal como se ilustra en la regresión de Z2 respecto a Z2 (o Z4 respecto a Z4), que


Figura 1.21: Apalancamiento (h_i)

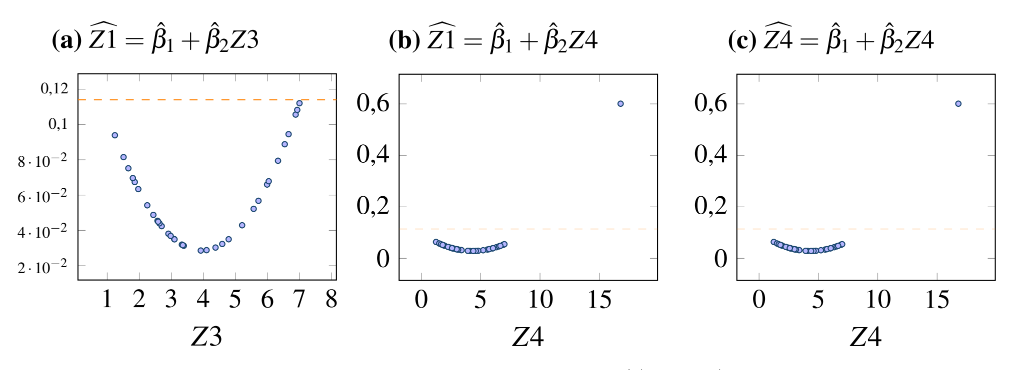
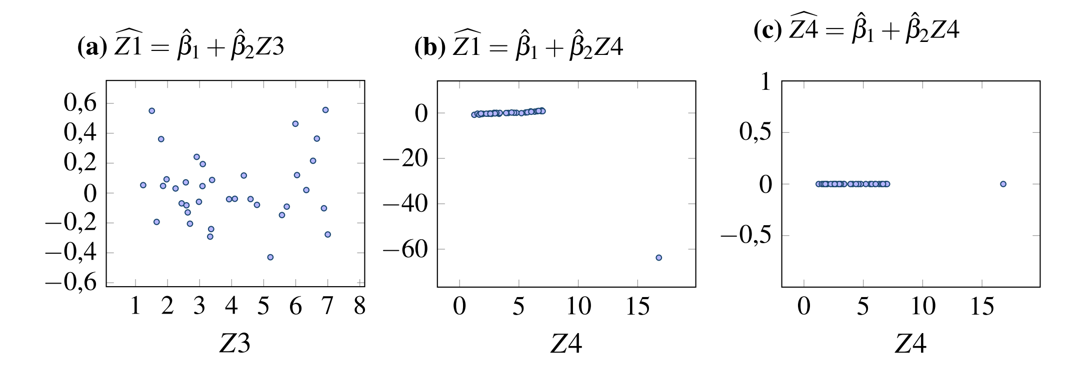


Figura 1.22: Influencia ($h_i \hat{u}_i / (1 - h_i)$)



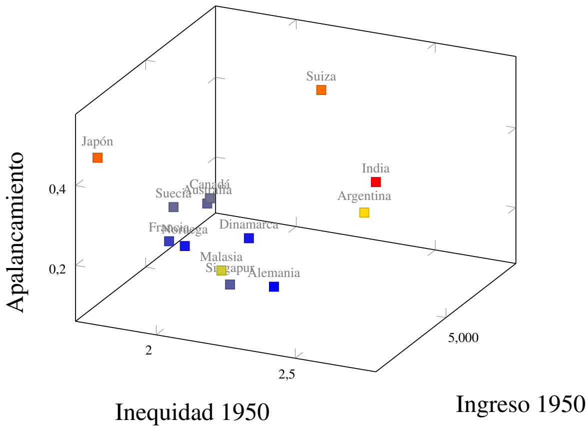
una observación tenga un alto nivel de apalancamiento no implica que tenga **influencia** (impacto sobre la recta estimada), la cual se mide con

$$\frac{h_i \hat{u}_i}{(1 - h_i)} \quad \text{o} \quad \frac{\hat{u}_i}{(1 - h_i)}. \tag{1.15}$$

En la figuras 1.21b y 1.21c se muestra cómo las regresiones respectivas tienen el mismo apalancamiento en sus n obseraciones (la línea punteada representa $2 \cdot k/n$ y muestra la presencia de una observación peligrosa) pero sólo en la regresión de Z1 respecto a Z4 el outlier está generando presión sobre la pendiente estimada.

Cuando se detecta un valor influyente, ¿qué se hace? Si queda claro que la observación es un outlier por motivos de, por ejemplo, un error de tipeo, entonces el dato debiera ser corregido o, en última instancia, eliminado.

Figura 1.20: Apalancamiento en M3



Hay quienes sucumben ante la tentación de eliminar las observaciones "molestosas", incluso cuando no se ha comprobado que presentan errores de medición. Esta no es

una práctica aceptable en la econometría, salvo que existan argumentos muy fuertes que respalden dicha decisión. Muchas veces existen mejoras al modelo que permiten mantener la totalidad de las observaciones. ¡Una observación no debe ser eliminada sólo porque no calza con el ideal del investigador!

GRETL: Influencia
El comando `leverage` tras una regresión `ols` presenta niveles de apalancamiento e influencia.

Figura 1.23: ¡Fuera mentiroso!



TO DO: Incorporar párrafo sobre control con dummies